CSC642- R

Flight Price Prediction

Nicholas Gdula

Advisor: Prof. Aguiar-Pulido

April 18, 2023

Table of Contents

1.0 Introduction and Background

The purpose of this project is to predict, using multiple regression methods, the price of flights that could be invaluable to customers seeking to find the most optimal time to purchase a

ticket. There are several factors that contribute to flight price, the dynamic pricing used by several companies makes picking the right time for a flight extremely difficult. The highly flexible practice of dynamic pricing is altering a services cost to reflect changing market conditions, particularly to mark up in times of high demand. This widely accepted industry business model has caused there to be a high demand for customers to buy tickets at the best time, or to at least have an idea of when it is best to book their ticket. The endogenous and exogenous factors that surround flight cost are highly complex which makes it arduous to build a one size fit all regression model without a significant data infrastructure or team for large data models. The goal of this model is to build a simple regression tool that could be used by customers that are booking flights for vacation or work travel, in other words this model could be utilized under normal day to day booking and seemingly would be less reliable when booking flights when there are external factors that could influence a particular price. These extrinsic factors would include booking a flight for festivals, natural disasters, concerts, or major sporting events. These examples represent specific times in which a particular flight out of a city may be dynamically priced higher due to demand, our model would not be able to predict without a more in-depth data set. Here, we will be looking at variables such airline, source city, departure times, number of layovers, arrival time, destination, flight class, days left, and duration. These variables represent replicable instances where a regression model may be able to make targeted predictions for flights not highly influenced by external factors. The target value is ultimately the flight price, which is measured in Rupees as we are dealing with companies headquartered in India.

Price predictors can play a key role in customer experience. With more and more companies competing for business, it is more important than ever to price your flights fairly. With the continued emergence of technology and the recent boom on data driven decisions the world

consumer can make more informed decisions than ever before. Any industry that provides a service there will almost always be a website or tool that consolidates the data so users can see competing prices. It is up to the company to then make their flight the most appealing by pricing it competitively for what service they are offering. Having a robust model to make these predictions of what flights should be at could not just benefit the customer but also the business deploying them. However, the vision for this research is to find a tool like that of Kelley blue book in the car industry where customers would be able to enter aspects of their flight and our tool could compare that to what our model believes to be the predicted price. Employing computational techniques could improve price prediction and could save the consumer from an otherwise uninformed decision. Within machine learning, researchers have aimed to tackle this problem from multiple angles. Traditional machine learning techniques have been implemented with success. Kelley Blue Book uses regression methods, although their algorithm is not stated. The industry of price prediction has been shown the most success using recurrent neural networks which are neural networks used for sequential data or XGBoost. The most recent publication on flight price prediction concluded that "no other research have included statistics from holidays, celebrations, stock market price fluctuations, depression, fuel price, and socioeconomic information to estimate the air transport market sector; nonetheless, there are numerous restrictions" (Biswas, 2022).

The source of our dataset was Kaggle, this site allows users to post work done with the dataset offered. These user submissions for the dataset used in this research use a variety of methods. Submissions differ in preprocessing techniques as well as the implementation of a variety of models. IBM describes preprocessing as a "broad set of techniques used to transform the data into a format that is more conducive to analysis and general data science techniques". The submissions using this dataset had several different ways of data preprocessing varying from using

what was possible from the given data, meaning no data preprocessing to hot encoding each possible column. This parity naturally resulted in the implementation of most common regression model however, the most common were linear regression, XGBoost, and random forest regression. The neural networks submitted had various architecture and were unique to each submission. Great strides have been made using recurrent neural networks to model time series data that has been shown to work in problems like the focus of this study. Most packages were submitted using python, the language of choice for this research will be R.

The most common themes of user submissions achieved an R^2 score of .92 and a MAE around 1500. The 1500 converted from Rupees to US Dollars is 18.27. Meaning most models had an average error of eighteen dollars and twenty-seven cents when comparing the actual and predicted price.

The data boom has shown the world that people want to be informed when it comes to making decisions with their money. This project seeks to facilitate that want for knowledge by accurately predicting what a flight should be priced at so consumers can make a decision that is best for them. Moreover, while the goal of this project is meant for flight price prediction similar methods could be applied to other service-based businesses.

2.0 Methods

2.1 Dataset

The dataset for this project has been taken from Kaggle:

https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction

The dataset contains three files with no missing values:

- Clean_Dataset.csv: this is the main training data. The number of rows is 300261 for all combinations of 11 predictors. The data was collected from Easemytrip.com using octoparse, a scraping tool that extracted data from the website. The data is over a fifty day period, February 11th to March 31st, 2022.
- business.csv: this is the business class flight data separated from the main dataset. (Will not be used, accounted for classes in regression model)
- economy.csv: this is the economy class flight data separated from the main dataset. (Will not be used, accounted for classes in regression model)

There are 9 variables that we will be considering for classifying the target and the explanation for each is given below:

- Airline : Airline Company
- Source_city: City in which flight is departing
- departure_time: Time in which flight leaves the source city
- stops: The number of layovers for flight
- arrival_time : Time landed in destination city
- destination_city: City in which flight is arriving
- class: Ticket class
- duration: Duration of the flight from the source city to destination city
- days_left: The amount of days remaining until flight departure
- price: Flight Price (target value)

Due to the large number of instances, the clean dataset was split into fifty percent training and fifty percent testing. The variables which give identity information are not considered for prediction. These variables are:

- Serial_number: flight identifier

- Flight: Flight code

In this section of the research paper, we describe the preprocessing steps taken in the R project to accurately predict flight prices. The dataset used in this project contains information about various flights, including their prices, airlines, source and destination cities, departure and arrival times, number of stops, and class of travel. The first step in the preprocessing of the data involves encoding the ordinal variables "stops" and "class" to allow for their use in the prediction model. This is done by converting the variables into factors and assigning them corresponding levels. The levels for "stops" are "zero", "one", and "two_or_more", while the levels for "class" are "Economy" and "Business". These variables are then transformed into integers by subtracting 1 from their respective values. Next, dummy variables are created for the categorical variables in the dataset. These variables include "airline", "source_city", "destination_city", "departure_time", and "arrival_time". Dummy variables are created by converting each categorical variable into a set of binary variables representing its categories. This is done using the 'model.matrix' function, which creates a matrix of dummy variables for the specified categorical variables. The resulting dummy variables are assigned appropriate column names by replacing spaces in the original variable names with periods. The original dataset is then merged with the matrix of dummy variables to form a new dataset. Finally, the original variables used to create the dummy variables are removed from the new dataset, resulting in a preprocessed dataset ready for use in the prediction model. The 'preprocessing' function takes the original dataset 'usedf' as input and returns the preprocessed dataset 'df_preprocessed'. This preprocessed dataframe has sixteen variables for the model

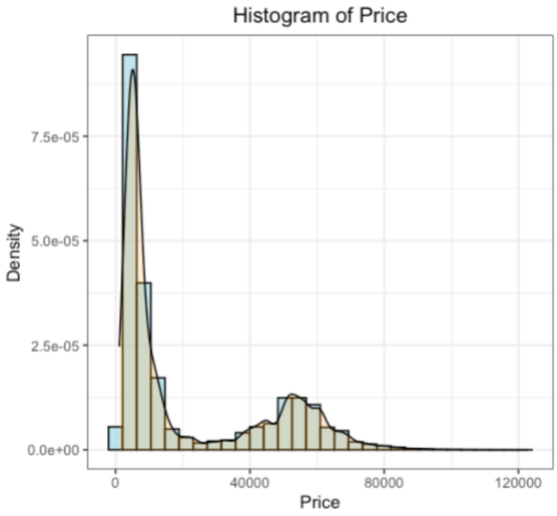
- stops
- class

- duration
- days_left
- airline_Indigo
- source_cityChennai
- source_cityDelhi
- source_cityHyderabad
- source_cityKolkata
- destination_cityHyderabad
- departure_timeEarly_Morning
- departure_timeEvening
- departure_timeLate_Night

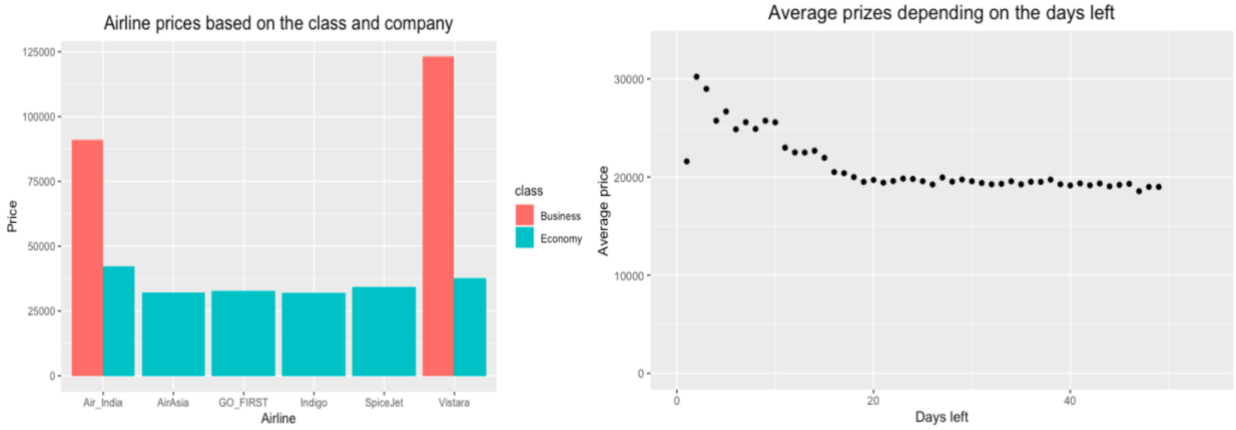
These preprocessing steps help to ensure that the dataset is in a suitable format for analysis and that the model is accurate in its predictions.

2.2 Experimental Data Analysis

The goal of experimental data analysis is to explore the data using visual plots to perform statistical analysis. The dataset provided was already clean so there were no missing values. Under normal circumstances, this would have been the first step in understanding the data. The next step is to examine the visual plots of our variables to better understand the influence they may have on our models as well as if there are any outliers. Creating both a boxplot and histogram of the flight prices one can see the general distribution of the data.



The flight data does not follow a normal distribution, this is likely because there are two classes of flights, business and economy. The bulk of the flights seem to be in the economy class with a smaller number of business class flights unbalancing the data. The plan was to manipulate the data to account for the categorical variable as its weight in the model may prove beneficial as it would have a large influence on the outcome. Considering creating a model for just business class and economy class flight prices may need to be considered. The next step was to visualize where the bulk of our data was coming from.



The above plots show us the prices of flights based on company and class as well as the average price of a flight depending on how many days are left before departure. The bar chart gives

insight on our previous suspicion that the bulk of flights are economy, this can be seen as there are only two companies that provided data on business class prices. The average price per company is also consistent with the thought that business class was the significantly more expensive when looking at the distribution. The price depending on days left seems to not have a large impact passed around the ten-day mark. This could be taken into consideration in regards to future pre-processing as the variable could be hot encoded to one if it is within ten days of departure and zero if greater than ten days out. There were visualizations for each chart that played a significant role in how the data was handled. The last plots that will be shown in the paper deal with the number of stops per company and the corresponding price.

These graphs were particularly important as they were able to show the variance in price based on the number of stops. Consideration was given to encode the variable to have either a layover or not. This would have eliminated the two_or_more instance which as one could see is consistently higher across all corporations. The thought behind keeping the zero, one, and two_or_more is that the weights in their respective models will be positively influenced by the presence, meaning having this ordinal variable remain as is could make predictions more precise. Other processes can be seen by referencing the code.

2.3 Evaluation

Eight regression algorithms were chosen for evaluation on our preprocessed dataset. These include Linear Regression, Bagging, Ridge Regression, Lasso, LDA, Decision Tree Regression, Support Vector Regression, and Random Forest Regression. These algorithms were chosen through evaluation of our dataset to see what has been popular in past work and facilitated what was learned in Professor Aguiar's 642-R class at the University of Miami. Each model will be trained on a tenfold cross validation method with a portion of the training set aside. Using a 10-fold cross-validation (CV) method and setting aside data for a validation set are both common techniques for assessing the performance of a machine learning model on a dataset. The primary reason for using a 10-fold CV method is to evaluate the generalization performance of the model. By splitting the data into 10 roughly equal parts, the model is trained on 9 of those parts and tested on the remaining 1 part. This process is repeated 10 times so that each part is used for testing once. By averaging the results across all 10 runs, we can obtain an estimate of the model's generalization performance on new data. Setting aside data for a validation set serves as a final check on the performance of the model before it is deployed in a production environment. The validation set is used to evaluate the model's performance on data that it has never seen before. This is important because it allows us to detect any overfitting or other issues that may have been missed during training and testing. For flight prediction, using a 10-fold CV method and setting aside data for a validation set is particularly important because the data has a temporal structure. Flights that occur close in time are likely to have similar characteristics, and using data from the future to predict flights from the past could lead to overfitting. By using a 10-fold CV method, we can ensure that the model is trained and tested on a representative sample of the data, while

setting aside a validation set can help us identify any issues with the model's ability to generalize to new data.

The choice of evaluation metric for a machine learning model is critical, as it determines how we measure the model's performance and, ultimately, its usefulness in solving the problem at hand. In the case of flight price prediction, there are several metrics that we could use to evaluate the performance of a model, such as mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). Of these metrics, RMSE is a particularly useful choice for flight price prediction because it considers the magnitude of the errors made by the model. Specifically, RMSE penalizes larger errors more heavily than smaller errors, which is important in the context of flight price prediction because even small errors in price predictions can have a significant impact on the profitability of an airline or in context of our goal, the satisfaction of a customer. Another advantage of using RMSE as an evaluation metric is that it is intuitive and easy to interpret. RMSE is expressed in the same units as the target variable (in this case, Rupees), which makes it easy to understand the magnitude of the errors made by the model in real-world terms. Furthermore, because RMSE is a commonly used metric in regression analysis, it allows us to compare the performance of our flight price prediction model with other models that use RMSE as their evaluation metric. It is worth noting that RMSE has some limitations as an evaluation metric. For example, it can be sensitive to outliers in the data, which may not be a desirable property in some applications. However, in the case of flight price prediction, we can mitigate this issue by using a robust regression model or by removing outliers from the training data. The team chose RMSE as the evaluation metric for the evaluation metric because it is a robust, intuitive, and widely used metric that takes into account the magnitude of errors and allows for easy interpretation and comparison with other models.

3.0 Results

Model	RMSE (Rupees)
Linear Regression	7112.69
Bagging	7112.41
Ridge Regression	0.48 ($0.0059 USD)
Lasso	16.47
LDA	Future Work
Decision Tree	6561.24
Support Vector Regression	Future Work
Random Forest Regression	Future Work

4.0 Conclusion

This research paper aimed to develop a flight price prediction model that could accurately forecast ticket prices for potential customers. Eight different models were trained and tested using a dataset consisting of sixteen variables. The linear regression model and bagging algorithm both produced similar RMSE values, indicating that they are suitable baseline models for this task. The decision tree model performed better than these baseline models, suggesting that it may be a useful model to consider in the future. However, it is worth noting that decision trees can be prone to overfitting and may require further tuning to generalize better. The ridge regression and lasso models were able to achieve remarkably low RMSE values by tuning the hyperparameter lambda. These models were chosen for hyperparameter tuning as they were the least computationally exhausting. They provide baseline assumptions that with hyperparameter tuning the precision of the models can excel. The Ridge Regression model's RMSE of 0.48 is

shockingly close as 0.48 Rupees converted to USD is $0.0059. Ridge regression may perform poorly if the assumptions of the model are not met, and lasso may struggle with datasets that have many features with weak effects. Overall, the results of this study demonstrate the potential for machine learning models to predict flight prices accurately. However, future research should aim to refine these models further and explore other variables that may affect ticket prices

4.1 Future Work

The LDA, SVM, and random forest models were introduced as potential models to consider in the future. These models have not yet been trained on the dataset, so we cannot draw any conclusions about their performance. However, it is worth noting that these models have hyperparameters that can be tuned to improve their performance. Energy should be given to tune every hyperparameter for each model to optimize results. Given appropriate resources the most intriguing research for flight price prediction could come in the form of a recurrent neural network. These networks have been shown to perform great on sequential and times series data. However, they are computationally expensive and need appropriate resources. The preliminary results of our research show great promise and should be investigated further.

5.0 References

Biswas, Prithviraj. “Flight Price Prediction: A Case Study.” International Journal for Research in Applied Science and Engineering Technology, vol. 10, no. 6, 2022, pp. 89–94., https://doi.org/10.22214/ijraset.2022.43666.

Groves, William, and Maria Gini. “On Optimizing Airline Ticket Purchase Timing.” ACM Transactions on Intelligent Systems and Technology, vol. 7, no. 1, 2015, pp. 1–28., https://doi.org/10.1145/2733384.

Gupta, Chitranjan Kumar, et al. “Dynamic Flight Price Prediction Using Machine Learning Algorithms.” 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2022, https://doi.org/10.1109/icac3n56670.2022.10074448.

Jansen, Tim. “A Linear Quantile Mixed Regression Model for Prediction of Airline Ticket Prices.” Radboud University , 3 Apr. 2013.